United Nations ECE/CES/2015/26

Economic and Social Council

Distr.: General
31 March 2015

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-third plenary session
Geneva, 15-17 June 2015
Item 4 of the provisional agenda
Modernization of statistical production and services and managing for efficiency

From domain-oriented to process-oriented production

Note by the Statistical Office of the Republic of Slovenia

Summary

The Statistical Office of the Republic of Slovenia began to modernize statistical production some years ago by launching a large infrastructure project. The project was based on the "Generic Statistical Business Process Model" and a few cornerstone concepts which should assure development of generalised but still flexible enough solutions to cover the needs of most of the statistical surveys. Instead of creating one overall general solution for the entire statistical process, the office chose a step-by-step approach. Generalised solutions for the specific parts of the process, for example data editing, aggregation, standard estimation, will gradually be developed and introduced into statistical production. The paper describes the main concepts of the development project, presents the results achieved so far and shares the plans for further development.

The paper is presented for discussion to the first session of the Conference of European Statisticians' seminar "Modernisation of statistical production and services and managing for efficiency".

GE.15-06706 (E)

1506706

Please recycle

I. Introduction

1. Implementation of statistical surveys is a demanding, time consuming and consequently quite expensive task. This is especially true in the case of surveys from the field of official statistics where the reliability of the statistical outputs is of special importance. It is undeniable that most of the official statistical surveys target very complex phenomena that are not easy to observe which makes the job of official statisticians even more demanding. With the constant pressure for budget cuts on the other hand, the official statisticians are hence more and more facing the challenge of producing the statistics of high quality with the significantly reduced resources.

2. The rapid development of information technology (IT) in recent years is offering possibilities to redesign the survey process in an entirely new and more effective way. It can be quite clearly observed that lately most national statistical offices have put a lot of effort into redesign activities at the institutional level. Terms such as re-organisation, modernisation, new architecture etc. are very frequent terms in the national statistical offices' strategic documents. The final goal that is usually stated in these documents is to use the wide range of newly developed information technology tools and applications to make the whole production cycle less burdensome and less expensive.

3. The Statistical Office of the Republic of Slovenia (hereinafter SORS) is no exception in this regard. The efforts to build a new modernised system for data processing already go several years back. Approximately ten years ago SORS started to plan modernisation of its processes with the aim to build fully integrated systems for the statistical production. The implementation of such a system did not work out as planned, so the approach gradually changed. SORS turned toward the development of several independent, generic IT solutions which would then for each particular survey be linked into the specific statistical process.

4. SORS is now carrying out a large infrastructure project with the aim to put the modernisation ideas into practice as part of the regular statistical production. Generalised solutions for the specific parts of the process, for example data editing, aggregation and standard estimation, will gradually be developed and introduced into the process of carrying out statistical surveys. The paper describes the main concepts of the development project, presents the results achieved so far and sketches the plans for further developments.

II. From fully integrated toward the modular solution

5. In 2007, SORS started a large project which aimed at creating a complex, fully integrated statistical production system that would be able to cover the whole production cycle for a vast majority of statistical surveys. This would, among other things, include a generic, metadata driven system for the whole cycle of the statistical data processing. As regards this generic system, at the end of the project SORS came to the conclusion that the results only partially met the expectations.

6. The project failed in its goal to create a fully integrated system which would link together all the particular tools that were developed throughout the project. The question that was consequently raised after the analyses of the project results was: Are such fully integrated systems really the best development orientation? In other words, do such integrated systems really assure the modernisation towards the flexible and (cost) effective statistical production system? All the discussions and considerations in the end resulted in the decision to give up the idea of a fully

integrated system and instead aim at slightly different solutions which still keep some important features of the "old system" (also called the metadata driven approach). The idea was to use a certain degree of disintegration which would contribute to larger flexibility while maintaining some generality of the system.

7. The main change in the approach was that SORS decided to break the statistical process into a set of smaller sub-processes and started to develop modular solutions for each of these sub-processes. These modular solutions should be designed in a way that they enable easy and flexible linking of inputs and outputs of the individual components to the whole statistical process. These components (also called the building blocks) should provide the generic software solutions for certain parts of the statistical production process and should be designed in a way that they can act independently. The main features of these building blocks could be summarized as follows:

(a) They are designed on the basis of harmonized, transparent and widely accepted methodological principles which have been determined before the actual creation of the particular building block;

(b) They should be opened to such extent that these building blocks can be plugged into different databases in different environments (e.g. Oracle, SAS) as long as the databases follow some basic rules for the organization of data;

(c) They are designed as fully metadata driven systems, meaning that information which determines the parameters for the execution of the processing for a specific survey and a specific reference period should be provided outside the core computer code. No information referring to specific survey execution should be incorporated into the general program code but should be provided by the subject-matter personnel through the special metadata tables;

(d) The process metadata can also be provided in different databases for each survey in different environments, but each of these (metadata) databases must follow strict rules regarding its structure (tables and variables).

8. The generic modular solutions were developed in two steps. In 2007-2010 the first version of the new generalised system was built. The focus of this initial development was to develop generic SAS based programs, which would be able to execute a particular part of the process for different surveys by not changing the SAS program itself, but just by adjusting the process metadata (process rules) which are the input into the general program. What was most important here was the significant change in the way the particular part of the statistical process was now executed.

9. Before introducing the new system, SORS applied "classical stove-pipe oriented production", where production tools and solutions were "survey-dependent". How the statistical processes (e.g. editing and imputations, sampling error estimation, statistical disclosure control and tabulation) were organized very much depended on the survey team in charge of the survey implementation. The survey-dependent approach was especially outstanding when the development of the software solutions was concerned. Software solutions were developed mostly ad-hoc for the needs of the particular survey. IT specialists used the "open instructions" of statisticians to develop the software solution and then (if needed) adjusted it for each particular implementation need when carrying out the survey. Such a system demanded a large amount of IT work at the development stage, and made the maintenance of these software solutions a very demanding job.

10. The main breakthrough that the new solutions presented was the fact that now the software solution for the particular part of the process was created only once and was then adjusted for particular implementations only through the process

metadata. These metadata are provided to the system by the survey statisticians which are now the managers of the whole system. IT specialists and general methodologists are only in the role to provide support when trouble occurs.

11. The first general solution was very flexible and open. The general programs could be plugged into different database environments for micro-data and for process metadata, if these data and metadata were organised according to the predefined rules. Such a highly generalized and open system is surely highly flexible and provides a suitable tool for building up a statistical process. However, there are also some quite obvious shortcomings of such an open system. These shortcomings are mostly connected to the process metadata management procedures. As indicated previously, the database of process metadata has a strictly determined structure, but it can for each particular survey be placed in different databases and even in different environments (e.g. Oracle, MS Access, SAS). In fact for most of the surveys the process metadata were stored inside MS Access databases. The reason for this was mainly the fact that subject-matter specialists, who are predominantly in charge of managing these metadata, prefer this environment due to its simplicity and user friendliness.

12. The problem with such a scattered system of process metadata is that it is impossible to create an effective general application for managing and controlling the inserted metadata. As it was pointed out in the analyses after the first period of the usage of the disintegrated system, the most problematic part was the significant number of errors in the process metadata. Since the fields for inserting rules are at the moment entirely open fields, most of these errors concerned the syntax of the rules (e.g. bracket errors) or errors in consistency between rules and variables.

13. To enable the creation of a better system for process metadata management and navigation, SORS decided to perform a certain degree of re-integration of the whole system. The aim of this re-integration is certainly not to build again the fully integrated system as initially designed, but to re-integrate only to such a level which would on one hand enable creation of the general management tool and on the other hand would keep the high flexibility of the system. The following re-integration actions were decided to be carried out:

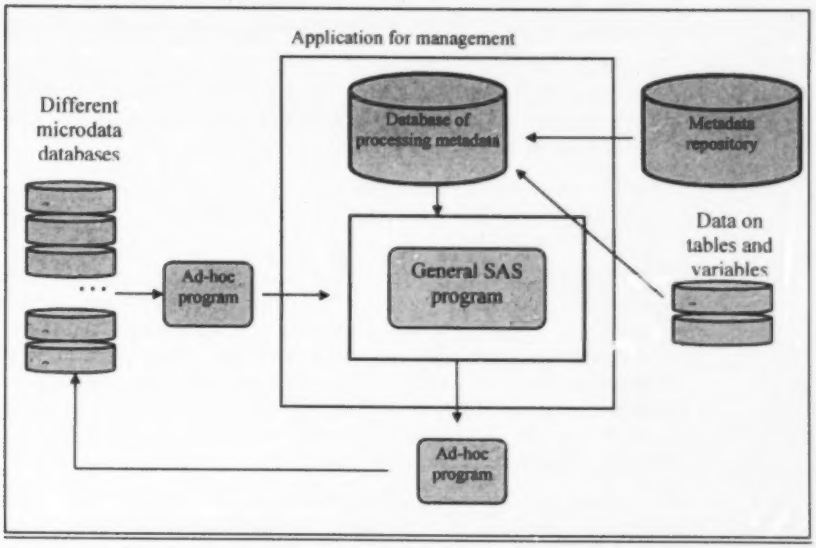
(a) To build one single, unique database of process metadata. This database would be created in Oracle and managed by the .NET application, which would enable user friendly management of the process metadata;

(b) To connect the system with the metadata repository, where the data on surveys and survey instances are stored;

(c) To enable the management application access to data on variables in incoming tables. These data can enter the application through two different channels: from the register of variables, if the survey has registered its variables there, or from some other location where the data tables structures are stored.

14. A simplified schematic presentation of the planned re-integrated system is presented in the following figure:

Figure 1
Schematic presentation of the reintegrated system



15. Development of this new system is the main subject of the large infrastructure project, mentioned at the beginning. SORS has already successfully finished the part dealing with the data validation, automated data corrections and data imputations and we are now gradually introducing this part of the application into regular production. The development of the building blocks for the data aggregation, disclosure control, standard error estimation and tabulation is also quite far. SORS is planning to start to introduce these modules into regular production at the end of this year.

III. Challenges when a totally new system is introduced

16. When the introduction of the new application started (with the very new approach) in the regular statistical production, SORS regularly collected feedback from statisticians that were faced with the new way of data processing. The main advantages and main drawbacks of the new approach, as perceived by them, can be summarised as follows.

Main advantages:

- The subject-matter personnel are much more independent of the IT department which was previously in charge of technical execution of the processes.
- The rules for the data processing can very quickly be changed through the centralised system of process metadata. This makes the whole data processing cycle much more flexible.
- Since the user can run the procedures several times in short time, it is now easier to check the feasibility of different methods for data processing.

Main drawbacks:

- In the process of the insertion of metadata there is a high risk of syntax errors. As the consequence, the application cannot be executed or is executed with the wrong parameterization.
- The subject-matter statisticians need to learn some new skills, which is sometimes a problem in the reality of the very burdensome statistical production.
- If an error occurs during the execution of the procedure, the technical staff must be contacted and if they are not available, the process execution can stop for some time.

17. Development of a totally new system for statistical production is certainly a big step forward for SORS. We firmly believe that the project outcomes will help us to build a new, modernised system for the management of statistical data processing in different statistical systems. Movement from stove-pipe to centralized methodological and IT solutions will be the final goal. The central point of the renovated system is the metadata driven application which is on one hand flexible in the sense that it can be plugged in different micro-data environments, while on the other hand it introduces a centralised management of the process metadata.

18. Introduction of such generic, metadata driven (MDD) application for data processing unavoidably introduces certain changes also at the general, institutional level when design and implementation of the statistical surveys are concerned. Based on the experiences gained so far, the main changes can be summarised as follows:

(a) There is essentially different distribution of work between subject-matter specialists, general methodologists and IT experts. With the old system, each subject-matter statistician had his or her "own programmer" and his or her "own general methodologists", who used the specific instructions of the subject-matter specialist to design and implement ad-hoc processes for a certain survey. Now the general methodologists and IT experts act only as the "support team" in case certain error in the application occurs or the process does not provide the expected results. This means that subject-matter specialists are now much more independent of the IT department and the general methodology department;

(b) Change in the role of subject-matter statisticians in the statistical process also changed expectations of their skills and capabilities. It used to be expected that they have a very deep knowledge of the subject-matter and that they are capable of providing the written instructions (in open form) for implementation of certain parts of the process (e.g. imputation, aggregation). Now they need to be trained and educated to be able to write these rules themselves already in the form of mathematical-computer language;

(c) The whole organization of work of the IT department and the general methodology department will have to be changed from domain oriented to process oriented. This re-organisation means a significantly different general view to the institution's organisation and distribution of work and is, therefore, quite a challenge for the statistical organisation. SORS is at the first stages of dealing with this challenge;

(d) The above described re-direction from (specific) domain oriented to (general) process oriented production will have to be realized also at the level of IT and methodology experts. Developing and supporting such generic applications requires experts capable of operating at a much more general level, considering the

execution of a certain survey just as one of the realisations of the general statistical process.

References

Dolenc, D., Krek, M., Seljak, R. (2011), "Editing Process in the Case of Slovenian Register- based Census", paper presented at the UNECE Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9-11 May 2011;

Seljak, R. (2009), "Integrated statistical systems and their flexibility – How to find the balance?", presented at the NTTS conference, Brussels, Belgium, 5-7 March, 2013;

Seljak, R., Blazic, P. (2011), "Sampling error estimation – SORS practice", Presented at the 2nd European Establishment Statistics Workshop, Neuchatel, Switzerland, 12-14 September, 2011;

Seljak, R. (2014), "Metadata driven application for data processing – from local toward global solution", paper presented at the UNECE Work Session on Statistical Data Editing, Paris, France, 28-30 April 2014.